DOCUMENT RESUME

ED 253 548 TM 850 085

AUTHOR	Dong, Hei-Ki; And Others
TITLE	An Empirical Investigation of Sample-Free Calibration Claim of the Rasch Model. Technical Report.
INSTITUTION	Ball Foundation, Glen Ellyn, IL.
REPORT NO	BF-TR-11
PUB DATE	Aug 83
NOTE	20p.; For a related study, see TM 850 087.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Ability; *Age Differences; Aptitude Tests; *Difficulty Level; *Goodness of Fit; Grade 9; Grade 12; Graduate Surveys; High Schools; *Latent Trait Theory; Mathematical Models; Post Secondary Education; *Test Items; Vocational Aptitude
IDENTIFIERS	Ball Aptitude Battery; Rasch Model

ABSTRACT

This paper examines the sample-free test calibration claim, derived from Rasch's structural model for test items, by comparing item difficulty and log ability estimates between different samples of ability levels. Three tests from the Ball Aptitude Battery--Inductive Reasoning, Paper Folding; and Vocabulary--were administered to three samples of different ability levels: (1) 353 high school freshmen; (2) 112 high school seniors; and (3) the same 112 high school seniors four years later as young adults. The Rasch model fit all three tests reasonably well in all three samples, and the claim of sample-free test calibration was supported in that the Rasch item difficulty and person ability parameters were relatively independent of the ability level of the calibrating samples. (Author/BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED253548

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Dong

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Technical Report No. 11

AN EMPERICAL INVESTIGATION OF SAMPLE-FREE CALIBRATION CLAIM OF THE RASCH MODEL

Hei-Ki Dong, Stephen M. Colarelli, and Yong H. Sung

The Ball Foundation

Elizabeth Rengel

University of Minnesota

August, 1983

The Ball Foundation
800 Roosevelt Road
Building C, Suite 206
Glen Ellyn, Illinois 60137

Abstract

This paper examines the sample-free test calibration claim derived from Rasch's structural model for test items. Three tests from the Ball Aptitude Battery--Inductive Reasoning, Paper Folding, and Vocabulary--were administered to three samples of different ability levels: 353 high school freshmen, 112 high school seniors, and the same 112 high school seniors four years later as young adults. The Rasch model fit all three tests reasonably well in all three samples, and the claim of sample-free test calibration was supported in that the Rasch item difficulty and person ability parameters were relatively independent of the ability level of the calibrating samples.

An Empirical Investigation of Sample-Free Test Calibration Claim of the Rasch Model[1]

In recent years many psychometricians have focused their research efforts on latent trait models. Because they imply "specific objectivity" in measurement (Rasch, 1960; Wright, 1968, 1977), these models offer solutions to many traditionally difficult psychometric problems, such as equating, mastery testing, and adaptive or tailored testing (Lord, 1980; Wright, 1977). Specific objectivity refers to what Wright (1968, 1977) calls "person-free test calibration" and "item-free person measurement." Person-free test calibration refers to estimating item (and test) difficulty independent of the sample; item-free person measurement suggests that a person's ability on a latent trait can be estimated without depending on a fixed set of items.

The most tractable latent trait model is the one-parameter logistic model, or the Rasch model (Rasch, 1960, 1961, 1966a, 1966b). This model involves three test requirements and three assumptions. It requires that tests measure only one trait or psychological function and that tests be nonspeeded and dichotomously scored. The three assumptions of the model are: (1) an individual's response on a test item is a function of the individual's ability, a, and the difficulty of the item, d; (2) the relationship between a and d is multiplicative, or, if logs are used,

additive; and (3) given $\underline{a}$ and $\underline{d}$, all responses are stochastically independent. In mathematical terms, the three assumptions are as follows:

(1) $P[\underline{x}_{ij}=|i,j] = \underline{s}_{ij}/1+\underline{s}_{ij}$, ($\underline{s}_{ij}>0$; i=1,2,...,$\underline{N}$; j=1,2,...,$\underline{n}$), where

$P[x_{ij}=1|i,j]$ = the probability of a correct response to item j by person i at each situation;

$\underline{s}_{ij}$ = the situational parameter.

(2) $$\underline{s}_{ij} = \underline{a}_i/\underline{d}_j,$$

where

$\underline{a}_i$ = i-th person's ability;

$\underline{d}_j$ = item j's easiness or difficulty.

(3) all responses ($\underline{x}_{ij}$'s) given $\underline{a}_i$ and $\underline{d}_j$, are stochastically independent.

From these assumptions, claims of person-free test calibration and item free person measurement are derived. The purpose of the present paper is to empirically examine the robustness of the person-free test calibration property.

Several studies (Anderson, Kearney, & Everett, 1968; Ironson, Guion, & Ostrander, 1982; Slinde & Linn, 1979; Tinsley & Dawis, 1975; Wright, 1968) have examined the person-free test calibration property of the Rasch model by comparing either item difficulty

estimates and/or log ability estimates across different groups of ability levels in the same or different samples. A limitation of these studies is that they have used samples of similar ability levels. For example, Wright (1968) in testing the person-free test calibration claim of the model used top and lower thirds of a sample of 976 beginning law students on 48 LSAT items. He tested the claim by comparing the log ability estimates for the two subsamples. A similar procedure of sample splitting was used by Ironson, Guion, and Ostrander (1982). Others who have used different samples have used samples that differed only marginally in ability levels. Anderson, Kearney, and Everett (1968) examined the item difficulty and person ability estimates on an intelligence test for two relatively similar samples: Australian Army recruits and Royal Australian Navy recruits. Slinde and Linn (1979) compared item difficulty on standardized reading achievement tests with "groups of examinees not widely different in ability" (p. 438); specifically, the groups were fourth, fifth, and sixth grade students. All of these studies support the findings that log ability and item difficulty estimates are relatively independent of the samples on which they were based.

In order to provide a more robust test of the person-free test calibration claim, it is necessary to compare item difficulty and log ability estimates between different samples of disparate

ability levels. The authors are aware of only one study (Tinsley & Dawis, 1975) which examined the person-free test calibration property on different samples of substantially different ability. Tinsley and Dawis (1975) examined item difficulties and log abilities across four disparate samples: college students, high school students, civil service clerical employees, and clients of the Minnesota State Division of Vocational Rehabilitation. Number analogy, picture analogy, and symbol analogy tests were used. These tests differed in content, but were similar in terms of the operations required, namely, analogy operations. Tinsley and Dawis found that item difficulty and log ability estimates were not, by and large, dependent on calibrating samples.

Since the Tinsley and Dawis (1975) study appears to be the only research which has examined the person-free test calibration property on different samples of different ability, it is important to see if their findings can be replicated and generalized using other samples and different types of tests. The present study empirically examines Rasch item difficulty and log ability estimates when groups of examinees differing widely in ability are used in the calibration of three different types of tests.

Method

Subjects

Two groups of subjects participated in the present study. The first group consisted of 353 high school freshmen from nine Chicago area high schools. Forty-five percent were male, 33 percent were black, 63 percent were white, and 4 percent were hispanic. The second group was comprised of 112 subjects from Illinois and Texas, tested on two different occasions: first, as high school seniors in 1976-77, and second, four years later as young adults in a part of longitudinal follow-up study. Forty-one percent were male, 14 percent were black, 68 percent were white, and 18 percent were hispanic.

Instruments

Three tests were selected from the Ball Aptitude Battery (BAB; Sung & Dawis, 1981), a multiple aptitude battery: Inductive Reasoning (30 items), Paper Folding (Spatial Ability; 24 items), and Vocabulary (80 items). The BAB tests are intended to estimate aptitudes or potential for learning specific work-related skills. Reliability estimates and evidence for the concurrent and construct validity of the BAB have been presented by Dohm and Sung (1980) and Sung and Dawis (1981).

The tests used in the present study met the Rasch model's requirement of dichotomously scored test items and they represented

a diversity of aptitudes. The BAB, including the three tests involved in this study, was administered in small groups. The group testing sessions lasted approximately three hours.

Analyses

Data analyses were carried out using the BICAL program (Wright & Mead, 1977). A total of nine BICAL analyses provided item difficulty and person ability estimates for each of the three tests administered at the three times of testing.

The person-free test calibration claim of the Rasch model was tested by examining the stability of item difficulty and person ability estimates across three different samples. This was evaluated by computing Pearson's product moment correlations among three samples' item parameter estimates and ability parameter estimates.

Results

Fit of the Rasch Model

The percentage of items fitting each of the three BAB tests are presented in the last column of Table 1. In the high school senior and young adult samples, the Rasch model fit the items in each test very well; it was able to fit 95 percent to 100 percent of the items in all three tests. However, in the high school freshman sample, the model did not perform as well. Seventy-one

percent and 79 percent of the items on the Paper Folding and Vocabulary tests, respectively, fit the Rasch model.

Insert Table 1 about here

Item Diffulty Estimates

Intercorrelations of Rasch difficulty estimates obtained from the three samples for each of the three BAB tests are shown in the lower portions of the triangles in Table 1. Correlations ranged from a low of .88 for the Vocabulary test in the freshmen versus young adults comparison to a high of .97 for the Paper Folding test in the seniors versus young adults comparison.

Person Ability Estimates

Upper triangle portions of Table 1 show the intercorrelations of Rasch ability estimates based on the three samples for each of the three BAB tests. Correlations from all the comparisons investigated were all .99. Figure 1 graphically illustrates the relationship between log ability estimates for the Inductive Reasoning test in the freshmen versus young adults samples.

Insert Figure 1 about here

Discussion

The results of this research indicate that the data from all three samples fit the model reasonably well and give support to the person-free test calibration claim of the Rasch model.

The Rasch model did not, however, fit the data in the freshman sample quite as well as it did in the other two samples. One reason might be that the data did not strictly meet the model's requirement of undimensionality. Results from factor analyses of test data suggest that the factor structure of three tests derived from the freshman data were not as undimensional as the factor structure derived from the senior data (Rengel, 1982). This also may be due to the data not strictly meeting the model's requirements of no guessing and nonspeededness: guessing and speededness of the tests may have been more of an influence in the freshman data than in the others as indicated by higher standard errors of measurement. Consequently, the Rasch model might not be as adequate for lower-ability examinees. In such cases, higher order models, such as the three-parameter latent trait model, which takes into account guessing, might be more appropriate. Although the basic concept of latent trait models is the same, the choice of one-, two-, or three-parameter model is currently an area of controversy and still subject to more emipirical research.

In general, the correlations in Table 1 extend the findings of previous studies to show that Rasch difficulty estimates and person ability estimates are relatively invariant across widely different ability groups on different types of tests, even when sample sizes were relatively modest. Two points, however, should be mentioned regarding correlations in Table 1. First, correlations of item difficulty estimates involving the freshman data were all lower than those involving senior and young adult data. This may be because of the dependency between the senior and young adult data and/or because the freshman data did not fit the model as well as data from the other samples. And second, presumably the invariance of both item and person parameter estimates across all three samples would be increased if aberrant items were deleted.

References

Anderson, J., Kearney, G.E., & Everett, A. V. 1968). An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 21, 231-238.

Dohm, T.E., & Sung, Y. H. (1980). Differential, concurrent, and construct validities of the Ball aptitude tests (Technical Report No. 5). Glen Ellyn, IL: The Ball Foundation.

Ironson, G. H., Guion, R. H., & Ostrander, M. (1982). Adverse impact from a psychometric perspective. Journal of Applied Psychology, 67, 419-432.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the fourth Berkeley symposium on mathematical statistics, IV, 321-334. Berkeley: University of California Press.

Rasch, G. (1966a). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Rasch, G. (1966b). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), Readings in mathematical social science (pp. 89-108). Chicago: Social Research Associates.

Rengel, E. (1983). Generalization and expansion of the Ball Aptitude Battery to high school freshman: psychometric properties, level differences, structure differences, and norm development (High School Freshman Research Project, Working Paper 1). Glen Ellyn, IL: The Ball Foundation.

Slinde, J. A., & Linn, R. L. (1979). The Rasch model, objective measurement, equating, and robustness. Applied Psychological Measurement, 3, 437-452.

Sung, Y. H., & Dawis, R. V. (1981). Level and structure differences in selected abilities across race and sex groups. Journal of Applied Psychology, 66, 613-624.

Sung, Y. H., & Dawis, R. V. (1981). Technical manual: Ball Aptitude Battery. Glen Ellyn, IL: The Ball Foundation.

Tinsley, H. E. A., & Dawis, R. V. (1975). An investigation of the Rasch simple logistic model: sample-free item and test calibration. Educational and Psychological Measurement, 35, 325-339.

Wright, B. D. (1968). Sample-free test calibration and person measurement. Proceedings of 1967 invitational conference on testing problems. Princeton, NJ: Educational Testing Service.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D., & Mead, R. J. (1977). BICAL: calibrating items and scales with the Rasch model (Research Memorandum No. 23). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Footnote

[1]An earlier draft of this report was given at the annual American Psychological Association Convention in Anaheim, CA, August, 1983.

Table 1

Correlations of Rasch Item Diffulty (Lower Triangle) and Person Ability (Upper Triangle) Estimates Along with Percentages of Items Fitting the Model

Comparison Sample	Comparison Sample			% Fitting the Rasch Model*
	H.S. Freshmen	H.S. Seniors	Young Adults	
Inductive Reasoning Test				
H.S. Freshmen	---	.99	.99	100%
H.S. Seniors	.91	---	.99	97%
Young Adults	.86	.94	---	100%
Paper Folding Test				
H.S. Freshmen	---	.99	.99	71%
H.S. Seniors	.95	---	.99	100%
Young Adults	.94	.97	---	96%
Vocabulary Test				
H.S. Freshmen	---	.99	.99	79%
H.S. Seniors	.93	---	.99	95%
Young Adults	.88	.96	---	95%

*Percentages were computed by counting the number of items that exceeded BICAL's fit value of ±2.50. (Therefore, the probability level was approximately .01.)

Figure Caption

Figure 1. Plot of raw test scores against log ability estimates for the Inductive Reasoning test.

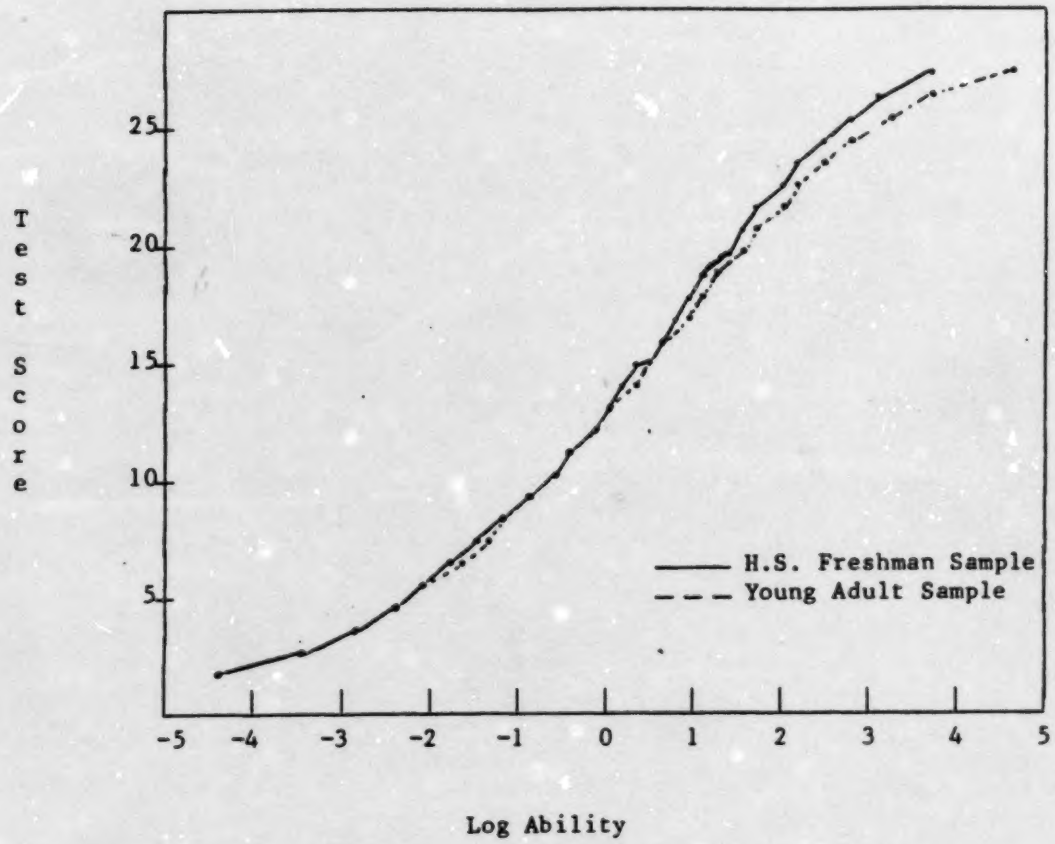

Figure 1. Plot of raw test scores against log ability estimates for the Inductive Reasoning test.